УДК 004.934

СИСТЕМА РАСПОЗНАВАНИЯ ГОЛОСОВЫХ КОМАНД

Ю.А. Зубань, И.В. Скляров
Сумский государственный университет

Исследования в области автоматического распознавания речи (Automated Speech Recognition, ASR) имеют различные цели. Упрощение интерфейса между пользователем и машиной для интеллектуального управления техническими объектами – одна из главных. Голосовое управление создает возможность организации природной связи между человеком и машиной. Разработка и исследование системы ASR с целью создания аппаратного комплекса управления техническими объектами при помощи голоса – основная цель данной работы.

ВВЕДЕНИЕ

На сегодняшний день благодаря развитию информационных и компьютерных технологий, программного обеспечения и внедрения их в разнообразные сферы человеческой деятельности все больше приводит к созданию новых подходов к "общению" с ЭВМ, в том числе основанных на речевом диалоге.

К работам в области распознавания речи относятся как работы, связанные с практически круглосуточным использованием распознавателей изолированных слов в промышленности и государственных учреждениях, так и научные исследования, ориентированные на создание универсальных распознавателей сложных предложений. Под распознаванием речи может пониматься преобразование речи в текст, распознавание и выполнение определенных команд, выделение из речи каких-либо параметров – все это в разных источниках может попасть под это определение.

Несмотря на очевидный прогресс в данной области исследований, автоматическое распознавание речи продолжает оставаться достаточно сложной проблемой. Существует ряд характерных особенностей речевых сигналов, в значительной степени затрудняющих решение этих проблем.

Все известные методы и приемы в распознавании речи не дают возможности явным образом определить, какой из видов анализа и какие параметры речевого сигнала могут дать наилучшие результаты. Вследствие этого работы по разработке алгоритма работы системы ASR сводятся к анализу уже существующих приложений ASR, а также перебору всевозможных методов параметризации сигнала и экспериментальному исследованию их эффективности.

ПОСТАНОВКА ЗАДАЧИ

Задачей данной работы являются разработка и отладка системы ASR с целью создания аппаратного комплекса управления техническими объектами при помощи голоса. Такая система может быть с успехом проинтегрирована в повседневные технические средства для облегчения взаимодействия человека и машины посредством голоса.

Разрабатываемая система должна иметь высокую вероятность правильного распознавания, высокий уровень производительности и надежности, обеспечивать работу в реальном времени, иметь достаточный набор голосовых команд и возможность работы с любым диктором.

ТЕХНОЛОГИЯ РАСПОЗНАВАНИЯ ИЗОЛИРОВАННЫХ ГОЛОСОВЫХ КОМАНД

Процедура распознавания голосовой команды состоит из двух основных этапов: этапа предварительной обработки, фильтрации и

выделения ключевых параметров речи и этапа сравнения входящей реализации команды с множеством предварительно созданных эталонов (шаблонов). Результатом распознавания будет эталон с минимальным расхождением входной реализации и шаблона словаря. Каждому шаблону ставится в соответствие определенное управляющее воздействие, с помощью которого и происходит управление техническим объектом.

Первым шагом в ASR является представление аналогового речевого сигнала в цифровом виде. В результате аналого–цифрового преобразования (АЦП) непрерывный сигнал переводится в ряд дискретных временных отсчетов, каждый из которых является числом. Таким образом, голосовой акустический сигнал, поступающий с микрофона при помощи АЦП, подвергается дискретизации и квантованию, т. е. производится оцифровка речевого сигнала. Главная цель процесса оцифровки заключается в получении цифровых данных с высоким отношением сигнал/шум, что необходимо для обеспечения высокой производительности приложений ASR. В описываемой системе сигнал дискретизируется с частотой 12 кГц и разрядностью 16 битов на каждый отсчет.

Для спектрального выравнивания речевого сигнала его следует пропустить через низкочастотный фильтр. В простейшем случае фильтрация осуществляется с использованием следующего соотношения:

$$v(n) = s(n) - \alpha \cdot s(n-1),$$

где $s(n)$ – исходный сигнал;

$v(n)$ – отфильтрованный сигнал;

α – параметр фильтрации.

Цель этого преобразования – снизить влияние локальных искажений на характеристические признаки, которые в дальнейшем будут использоваться для распознавания. Кроме того, слух человека является более чувствительным в регионе спектра выше 1 кГц, а данный фильтр усиливает именно эту область, помогая алгоритму спектрального анализа в выделении наиболее важных аспектов речевого спектра [7].

Считается, что на временном интервале порядка 8 – 20 мс форма голосового тракта существенно не меняется, поэтому характеристики речи на данном интервале времени остаются постоянными. Каждая команда разбивается на последовательность кадров (фреймов) длительностью 20 мс, что соответствует 240 отсчетам сигнала.

Следующим этапом описываемой системы является отделение полезного сигнала голосовой команды от фонового шума до и после ее реализации. Метод обнаружения моментов начала и окончания фразы также используется для уменьшения числа арифметических операций, так как дальнейшей обработке подлежат только те сегменты, в которых имеется речевой сигнал.

Предполагается, что первые 100 мс не содержат полезного сигнала, а содержат только шум. Следовательно, для детектирования начала голосовой команды необходимо сначала подсчитать энергию шума. Далее, кадр за кадром, подсчитывается энергия сигнала, и если она превышает шумовой порог – детектируется начало команды, – в противном случае фрейм удаляется. После обнаружения начала команды аналогично находится и ее окончание, с учетом того, что анализ идет в противоположном направлении. Данный алгоритм эффективен лишь при высоких соотношениях сигнал/шум, если же уровень шума велик, для детектирования согласных необходимо использовать совместно с энергией величину числа переходов сигнала через нулевой уровень [1].

Речевые сигналы характеризуются чрезвычайным разнообразием и изменчивостью. Существует множество различных факторов, которые

негативно сказываются на производительности системы ASR: настроение и эмоциональное состояние диктора, скорость произнесения фразы и отдельных ее элементов. При этом у диктора могут меняться темп речи, высота основного тона, ширина динамического диапазона и т.д. Все эти факторы, в свою очередь, негативно сказываются на всем процессе распознавания в целом. Значит, для того чтобы различать речевые сигналы, необходимо выделять такие параметры, которые были бы максимально инвариантны ко всем вариациям речи и несли в себе максимум фонетической информации о значении команды и минимум данных о самом дикторе, его эмоциональном и физическом состоянии.

Для формирования вектора параметров в системах ASR наиболее широко применяются методы спектрального анализа, так как наиболее важная информация скрыта именно в частотной области сигнала. В данной работе в качестве вектора параметров используются кепстральные коэффициенты, полученные с помощью преобразования Фурье. Для их вычисления спектр речевого сигнала приводится к тому состоянию, в котором он поступает на обработку механизмами человеческого уха (к мел–шкале), логарифмируется и подвергается обратному преобразованию Фурье. В результате получается кепстр (Mel–Frequency Cepstral Coefficient, MFCC), 12 первых коэффициентов которого содержат всю, имеющую существенное значение информацию об огибающей спектральной характеристике [2, 3, 5, 7, 8].

После вычисления необходимых оценок параметров их необходимо сравнить с предварительно записанными эталонами голосовых команд. Поскольку диктор не в состоянии повторить абсолютно точно в том же темпе одну и ту же фразу, нецелесообразно сравнивать кепстральные траектории, которые являются функциями времени. Эту трудность можно преодолеть путем нелинейного преобразования временного масштаба входных параметров для получения наиболее точного соответствия между эталоном и распознаваемой реализацией команды.

В разрабатываемой системе применяется метод на основе динамического программирования – динамическое искажение времени (Dynamic Time Warping, DTW) [1, 5, 6], прекрасно себя зарекомендовавший в ASR изолированных слов и задачах идентификации диктора по голосу. Техника динамического искажения используется для временного вытягивания и сокращения расстояния (времени) между эталонной траекторией и искаженным спектральным представлением неизвестной голосовой команды. Метод динамического искажения используют практически все коммерчески доступные системы распознавания, работающие с малым словарем, показывающие высокую точность распознавания.

Заключительным шагом в процессе распознавания команды являются вычисление некоторой полной меры различимости между входной реализацией и словами из библиотеки эталонов. Глобальная оценка расстояния между двумя векторами признаков состоит из локальных оценок и подсчитывается для каждого эталона словаря с помощью эвклидовой нормы [1]:

$$D(x, y) = \sqrt{\sum_i (x_i - y_i)^2} ,$$

где D(x,y) – эвклидова норма между входным вектором x и вектором–эталоном y;

x – входной вектор;

y – эталонный вектор словаря.

Из полученных глобальных оценок для каждого шаблона словаря выбирается оценка с минимальным значением эвклидовой нормы – это и

будет результат распознавания голосовой команды. Ей в соответствие ставится управляющее воздействие системы.

ПРАКТИЧЕСКИЕ РЕЗУЛЬТАТЫ

Для экспериментального исследования описанного алгоритма в среде MatLab 6.5 была реализована и исследована система распознавания изолированных голосовых команд с предварительным обучением и настройкой на диктора. В качестве анализируемых команд использовались речевые сигналы, содержащие слова русского языка. Словарь состоял из 15 специально подобранных команд средней длительностью 500 – 900 мс. Установлено ограничение на максимальную длину голосовой команды, которая не должна превышать 1,2 секунды.

Прежде чем приступить к работе с системой, ее необходимо обучить – создать индивидуальную базу эталонов для конкретного диктора. Процесс обучения является частью алгоритма распознавания, за исключением динамической коррекции траекторий и подсчета глобальных оценок различий. После вычисления MFCC вектор параметров сразу сохраняется в памяти ЭВМ и в дальнейшем используется системой как эталон.

В процессе работы с системой (как обучение, так и распознавание) к диктору предъявляются строгие требования на произношение: все слова должны произноситься с одинаковой громкостью, четко и монотонно, чтобы свести к минимуму влияние нежелательных эмоционально–акустических факторов и, следовательно, снизить вероятность ошибки.

При исследовании было установлено, что выбор распознаваемых слов очень важен. Чем сильнее слова в словаре фонетически отличаются, тем проще их распознавать, тем выше вероятность правильного решения и производительность системы ASR в целом. Это важное свойство необходимо учитывать при составлении словаря эталонов.

При соблюдении приведенных рекомендаций при работе с системой достигнут довольно высокий процент верных распознаваний (92–95%), что дает возможность использовать данный алгоритм при создании интеллектуальных систем управления техническими объектами.

ВЫВОДЫ

Для решения поставленной задачи надежного распознавания изолированных голосовых команд предлагаемый метод является наиболее оптимальным. Основные достоинства этого метода заключаются в простоте реализации, универсальности и достаточно высокой скорости вычислений по сравнению с другими методами распознавания. При помощи данного алгоритма можно успешно решить представляющую практический интерес задачу надежного распознавания отдельно произнесенных голосовых команд для интеллектуального управления техническими объектами.

Быстродействие современных ЭВМ открыло новые возможности в области разработки систем ASR: вычисление анализируемых параметров речи, сравнение с эталонами и динамическая коррекция, выполняемые с помощью быстродействующего цифрового сигнального процессора (Digital Signal Processor, DSP), обеспечит функционирование системы в реальном времени.

SUMMARY

There are different goals of research in area of Automated Speech Recognition. Machine interface simplification for smart technical objects management is one of main aims. A voice management gives a possibility to produce the natural contact between a man and machine. In this paper authors consider the real ASR system development for hardware complex of technical objects voice management.

СПИСОК ЛИТЕРАТУРЫ

1. Рабинер Л.Р., Шафер Р.В. Цифровая обработка речевых сигналов /Пер. с англ. – Москва: Радио и связь, 1981. – 496 с.
2. Бондарев В., Трёстер Г., Чернега В. / Цифровая обработка сигналов: методы и средства: Уч. пособие для вузов. – 2 – е изд. – Харьков: Конус, 2001. – 398 с.
3. Применение цифровой обработка сигналов /Под ред. Э.Оппенгейма. – Москва: Мир, 1980. – 545 с.
4. Рабинер Л., Гоулд Б. Теория и применение цифровой обработки сигналов. – Москва: Мир, 1978. – 835 с.
5. S. Furui. Digital Speech Processing, Synthesis and Recognition. Second Edition Revised and Expanded. – Tokyo: CRC, 2000. – 452 p.
6. J. Picone. – Modeling signal in speech recognition. – Proceedings if the IEEE – 1993.
7. S. Furi, M. Sondi. Advances in Speech Signal Processing. – New York: Marcel Dekker, inc., 1991. – 871p.

Ю.А. Зубань, канд. техн. наук
Сумский государственный университет

И.В. Скляров, студент
Сумский государственный университет

Поступила в редакцию 20 марта 2007 г.